Result Integrity Check for MapReduce Computation on Hybrid Clouds

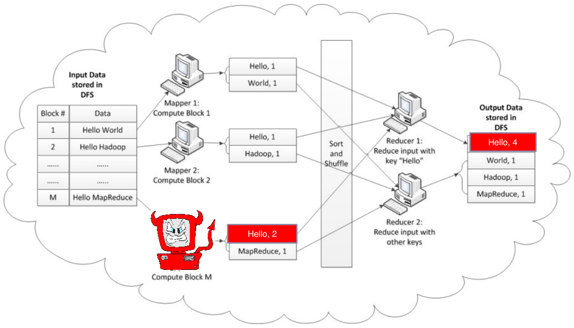
Yongzhi Wang*, Jinpeng Wei*, Mudhakar Srivatsa §

*** Florida International University**

§ IBM T.J. Watson Research Center

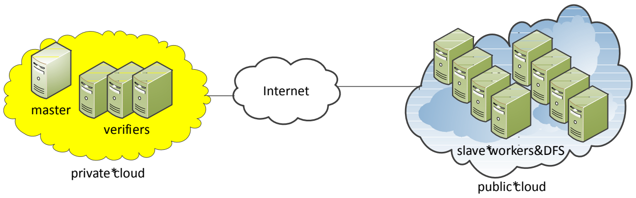
Motivation and Solution

- **Large-scale adoption of MapReduce computations on public clouds is hindered by the lack of trust on the participating virtual machines, because misbehaving worker nodes can compromise the integrity of the computation result. Therefore, a solution is needed to offer high result integrity while incurring a modest performance overhead.**



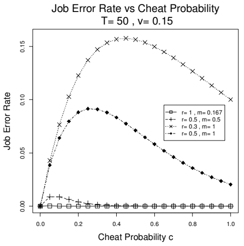
- **We propose a novel MapReduce framework, Cross Cloud MapReduce (CCMR), which overlays the MapReduce computation on top of a hybrid cloud: the master that is in control of the entire computation and guarantees result integrity runs on a private and trusted cloud, while normal workers run on a public cloud.**

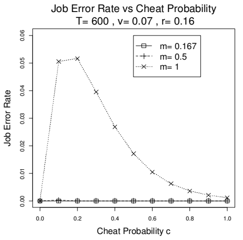
System Architecture



- The master and a small number of trusted workers (verifiers) are deployed to the trusted private cloud.
- Other slave workers and the DFS (Distribute File System) are deployed to the untrusted public cloud.
- Since task assignment and result checking are performed on the private cloud, which is controlled by the cloud user, "trust" is retained for the cloud user.

Theoretical Analysis (Accuracy)

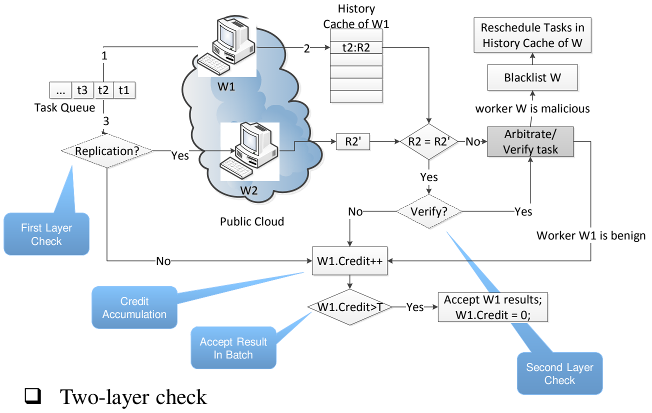




- r: replication probability
- v: verification probability
- T: credit threshold
- m: malicious worker ratio
- c: cheat probability

Reference: Yongzhi Wang, Jinpeng Wei, Mudhakar Srivatsa "Result Integrity Check for MapReduce Computation on Hybrid Clouds". The 6th IEEE International Conference on Cloud Computing (IEEE CLOUD 2013), June 27-July 2, 2013, Santa Clara, CA.

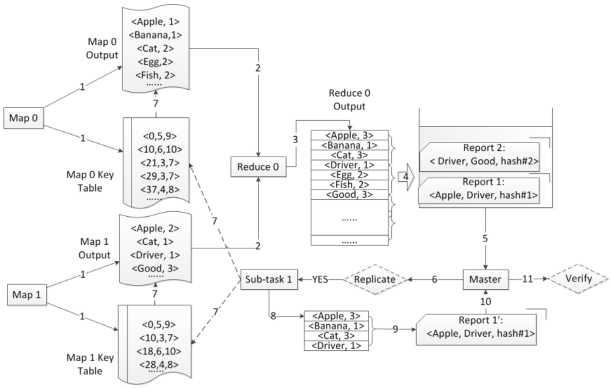
Map Phase Integrity Check



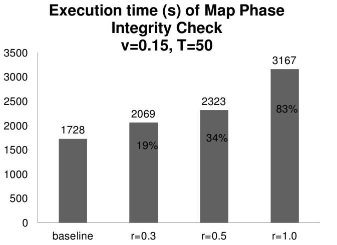
- ❑ Two-layer check
  - Probabilistically replicate tasks and check the consistency of returned task results (hash values)
  - Probabilistically verify tasks with consistent results (due to collusions)
- ❑ Credit-based Trust Management
  - Accumulate a worker's credit when it passes a Two-layer check
  - Accept a worker's task results in a batch when a worker achieves certain credit threshold

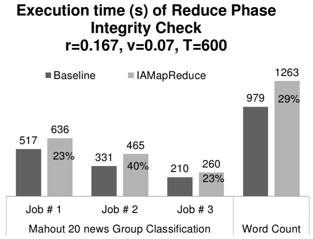
Reduce Phase Integrity Check

- Motivation: In some jobs, the number of reduce tasks is small while the number of processed records in each task is huge.
- Solution: Divide each reduce task into sub-tasks and apply two-layer check and credit based trust management to each sub-task.



Experimental Result (Performance)


Execution time (s) of Map Phase Integrity Check
v=0.15, T=50
baseline
r=0.3
r=0.5
r=1.0


Execution time (s) of Reduce Phase Integrity Check
r=0.167, v=0.07, T=600
Baseline
IAMapReduce
Job # 1
Job # 2
Job # 3
Mahout 20 news Group Classification
Word Count